

Strigil

A framework for data extraction

Peter Zvirinský
MFF CUNI

Project Strigil

- Represents an extendible, scalable and complete web scraping tool with RDF output.

Strigil Script

```
<scr:script xmlns:scr="http://sourceforge.net/projects/strigil/" version="1.0" type="HTML" pc: http://purl.org/procurement# type="HTML">
  <!-- root template -->
  <scr:call-template name="rootTemplate" type="http/GET">
    <scr:value-of text="http://www.isvzus.cz/cs/Searching/ContractWinners" />
  </scr:call-template>

  <scr:template name="rootTemplate" mime="text/html">
    <scr:call-template name="ContractDetail" type="http/GET">
      <scr:value-of select="div#SearchGrid div.t-grid-content table tr td a @href" />
    </scr:call-template>
  </scr:template>

  <scr:template name="ContractDetail" mime="text/html">
    <scr:onto-elem rel="pc:Contract" typeof="pc:Contract" about="">
      <scr:value-of select="input#FormItems_SpisCislo_IV_3_1 @value" property="pc:referenceNumber" />
      <scr:value-of select="textarea#FormItems_StrucnyPopis_II1_4" property="pc:description" />
    </scr:onto-elem>
  </scr:template>
</scr:script>
```

Web UI

STRIGIL

[Home](#) [Ontologies](#) [Scripts list](#) [Results](#) [Download constraints](#) [Logs](#) [Users list](#) [Log out](#)

[Script](#) [Templates](#)



Save

rootTemplate

rootTemplate

Variables & Parameters

Elements

 contractDetail
(CallTemplate)
 (ValueSelectType)

Element edit

Name: ...

Select: body div#main div#panel-left div#menu ul li.folder ul li


Text: ...

Var: ...

Regexp: ...

Replace: ...

Ontologies

**MINISTERSTVO
PRO MÍSTNÍ
ROZVOJ ČR**
VĚSTNÍK VEŘEJNÝCH ZAKÁZEK

Úvodní stránka
Aktuality
Vyhledávání
Podle více parametrů
Podle data uveřejnění
Podle evidenčních čísel
Podle názvu zakázky
Podle zadavatele
Vítězové veřejných zakázek
Seznam profilů zadavatelů
Seznam zrušených profilů
zadavatelů
Provést uveřejnění
Podání formuláře k uveřejnění
Číselníky a klasifikace
používané při uveřejňování

Nová vyhlášení

Evidenční číslo	Název zadavatele:	Předmět zadávacího postupu
7205011032748	ČEZ, a.s.	1(2)RH21,22,23\$102, 1(2)RH31,32,33\$102 - (ZSE 2008 306 ETE)
7203020032275	Český rozhlas	Úprava newsroomu Radiožurnálu v objektu Českého rozhlasu, Římská 13, Praha 2
7203011012804	Západočeský Plzeň	Kapilární elektroforéza s UV detekcí
7202011026904	Fakultní nemocnice Plzeň	Spotřební materiál pro dialýzu 2012

Noví vítězové zakázek

Evidenční číslo	Název dodavatele	Předmět zadávacího postupu
-----------------	------------------	----------------------------

Excel Addon

5rm_m33_p1_2012.xls [jen pro čtení] [režim kompatibility] - Microsoft Excel (Zkušební verze)

Soubor Domů Vložení Rozložení stránky Vzorce Data Revize Zobrazení Doplněk Team

save close save as panel open Strigil

H1

Veřejné zakázky detailně Statutární město Ústí n.L.

Odbor	Celková částka v Kč	Počet zakáz	Název dodavatele	Číslo zakáz	Popis zakázky	Částka v Kč	Čerpáno (včetně DPH)	Poptávaný subjekt
						53 882 380,67		Z.E.R.O.STAV INŽENÝRING a.s.
						0,00		Vodohospodářské stavby Teplice
						0,00		VIAMONT Development a.s.
Celkem za V. pásmo: nad 2 000 000	84 438 884,75	7						
0117 Odbor dopravy	9 825 418,37	3						
			INSKY, spo.s r.o.	VZ1171110006	Velkoplošná pokládka ABS II po zinním c	2 733 030,00	3 279 636,00	INSKY, spo.s r.o.

od 50 000 do 150 000 od 150 000 do 500 000 od 500 000 do 2 000 000 nad 2 000 000

RDF Data taskPane

```
<st0:script xmlns:st0="http://sourceforge.net/projects/strigil" type="EXCEL" id="UnL_001"
prefix="pc: http://purl.org/procurement/public-contracts#
owl: http://www.w3.org/2002/07/owl#
rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
gr1: http://purl.org/goodrelations/v1">
<st0:params>
<st0:param name="URI"></st0:param>
</st0:params>
<st0:meta status="draft" domain="" author="simon.dembinny@gmail.com" frequency="10d" max
<st0:call-template name="main" type="xls"></st0:call-template>
<st0:template name="name" mime="xls">
<!--Main template-->
<st0:ontoElem typeof="pc:Contract">
<
[pc:Contract 15011 tc:hdr=1číslo zakázky] td:group tr">
<st0:ontoElem
Description for st0:ontoElem></st0:ontoElem>
<st0:value-of
<st0:concatenate">
am>
<st0:switch
<st0:call-template
am>
```

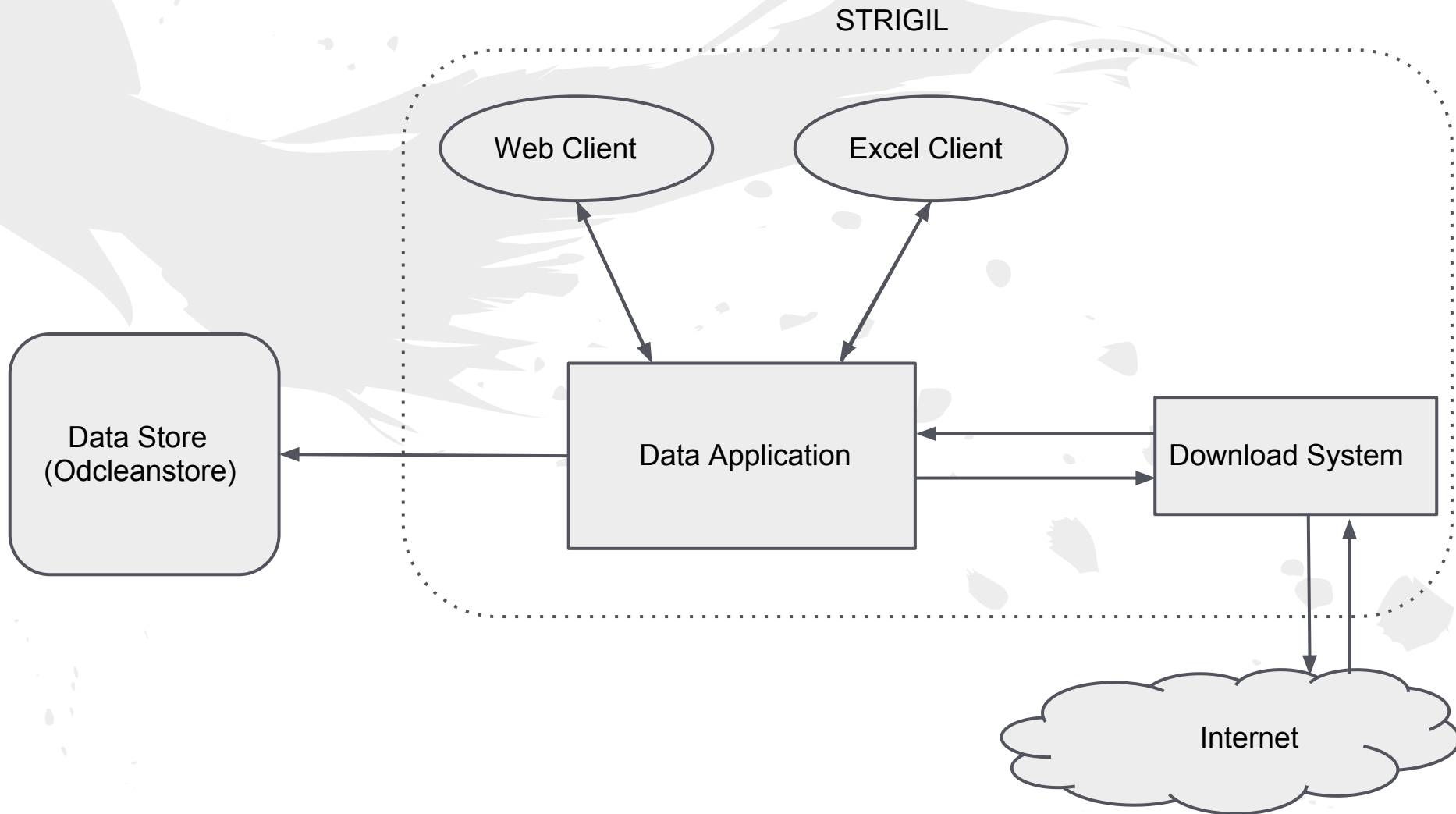
OWLDotNet.Ontologies.goodrelations_v1.owl

OWLDotNet.Ontologies.annotations_vann.owl

- #Wednesday
- #hasNext
- rdfs.range → #DayOfWeek
- #hasPrevious
- #DeliveryMethod
- #LocationOfSalesOrServiceProvisioning
- #N-Ary-Relations
- #Offering
- #OpeningHoursSpecification
- #PaymentMethod
- #PriceSpecification
- #QualitativeValue
- #QuantitativeValue
- #TypeAndQuantityNode
- #hasBusinessFunction
- rdfs.range → #BusinessFunction

- wks[name="List1"] (1)
- tr (1)
- wks[name="List1"] (1)
- tc[hdr="Hello"] (1)

Architecture



Results

- finished scraping on various public data sources
 - <http://www.isvzus.cz/>
 - <http://www.mpsv.cz>
 - <http://www.ted.europa.eu>
 - <http://www.ezak.cz/>
- integration with ODCleanStore

Project Info:

- web: <http://strigil.sourceforge.net/>
- release: 29. November 2012

